# Marginalization using the metric of the likelihood

R. PREUSS* and V. DOSE

*Max-Planck-Institut für Plasmaphysik, EURATOM Association*

*Boltzmannstr. 2, D-85748 Garching b. München, Germany*

(Dated: January 28, 2003)



## Abstract

Although the likelihood function is normalizeable with respect to the data there is no guarantee that the same holds with respect to the model parameters. This may lead to singularities in the expectation value integral of these parameters, especially if the prior information is not sufficient to take care of finite integral values. However, the problem may be solved by obeying the correct Riemannian metric imposed by the likelihood. This will be demonstrated for the example of the electron temperature evaluation in hydrogen plasmas.

*Electronic address: preuss@ipp.mpg.de

## I. INTRODUCTION

Given data $\vec{d}$, a linear parameter $c$ and some function $\vec{f}$ meant to explain the data, we have

$$\vec{d} = c \cdot \vec{f}(T) + \vec{\varepsilon} \qquad . \tag{1}$$

The vectors shall have dimension $N$ according to the number of quantities measured. Due to the measurement process the data is corrupted by noise, where $\langle \varepsilon \rangle = 0$ and $\langle \varepsilon^2 \rangle = \sigma^2$. Then by the principle of Maximum Entropy the likelihood function reads

$$p(D|c, \sigma, \vec{f}, I) \propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_i [d_i - c f_i]^2 \right\} \qquad , \tag{2}$$

which is clearly normalizeable for the data $\vec{d}$ and bound for every parameter showing up as a functional dependency in $f$. The situation may change when we are looking for the expectation value of some parameter of $f$, let say $f = f(T)$. Then we need to evaluate the posterior of $T$ with

$$\langle T \rangle \propto \int T \; p(\mathrm{d}T|D, I) \qquad . \tag{3}$$

In order to connect the unknown posterior to the known likelihood we marginalize over all the parameters which enter the problem, that is in our problem $c$ and $\sigma$:

$$p(\mathrm{d}T|D, I) = \int_c \int_\sigma p(\mathrm{d}T, \mathrm{d}c, \mathrm{d}\sigma|D, I) \qquad , \tag{4}$$

and make use of Bayes theorem:

$$p(T, c, \sigma|D, I) \propto p(D|T, c, \sigma, I) \; p(T, c, \sigma|I) \qquad . \tag{5}$$

Commonly, the infinitesimal elements in equation (4) are identified with

$$p(\mathrm{d}T, \mathrm{d}c, \mathrm{d}\sigma|D, I) = p(T, c, \sigma|D, I) \; \mathrm{d}T \; \mathrm{d}c \; \mathrm{d}\sigma \qquad . \tag{6}$$

In mathematical terms this would mean that the probability functions live in euclidean space. They do not.

## II. RIEMANNIAN METRIC

Parameterizations correspond to choices of coordinate systems. The problem to be solved has to be invariant against reparametrizations [1], i.e. in the space of the probability functions

one has to get the same answer no matter what parameters were chosen to describe a model. Therefore one is in need of a length measure $\mu$ which takes care of defining a distance between different elements of this probability function space. This task is done by applying differential geometry to statistical models, an approach which was baptized 'information geometry' by S. Amari [2]. Eq. (6) then reads correctly

$$p(\mathrm{d}T, \mathrm{d}c, \mathrm{d}\sigma|D, I) = p(T, c, \sigma|D, I)\ \mu(\mathrm{d}T, \mathrm{d}c, \mathrm{d}\sigma) \qquad . \tag{7}$$

$\mu(\mathrm{d}\vec{\theta}) = \mu(\vec{\theta})\mathrm{d}\vec{\theta}$ is the natural Riemannian metric on a regular model (in our case the model is parameterized by $\vec{\theta} = (T, c, \sigma)$). It results from second variations of the entropy [2, 3] and is given by

$$\mu(\mathrm{d}\vec{\theta}) = \sqrt{\det \boldsymbol{g}(\vec{\theta})}\mathrm{d}\vec{\theta} \tag{8}$$

where $\boldsymbol{g}$ is the Fisher information matrix:

$$g_{ij} = -\left\langle \frac{\partial^2 \log p(D|\vec{\theta}, I)}{\partial\theta_i \partial\theta_j} \right\rangle \qquad . \tag{9}$$

For the above likelihood the metric reads explicitly

$$\mu(\sigma, c, T) \propto \frac{c}{\sigma^3}\sqrt{\left[\sum_i f_i^2\right]\left[\sum_i \left(\frac{\partial f_i}{\partial T}\right)^2\right] - \left[\sum_i f_i \frac{\partial f_i}{\partial T}\right]^2} \qquad . \tag{10}$$

Notice that this approach is based on the assumption that the hypothesis space of the likelihood defines the metric to be calculated in. This may not be the case if some prior information was already used during data acquisition, e.g. the experimentalist uses his expert knowledge in separating 'correct' data from the rest. The latter instantly rules out certain parts of all possible realizations of the likelihood function and results in a different hypothesis space.

## III. SIMPLE EXAMPLE

First we want to demonstrate the relevance of using the correct metric with a simple example which already has all the features of the real world problem further down.

$$f_i(T) = T \cdot (T + x_i)^{-1} x_i \qquad , \tag{11}$$

where the notation in $i$ corresponds to the data points $d_i$. For simplification let us assume that the variance $\sigma^2$ is known and we only have to marginalize over $c$ in order to get the

posterior. What happens if we do not use the Riemannian metric? Then the marginalization integral over $c$ reads

$$p(T|D,I) \propto \int \mathrm{d}c \; p(D|T,c,I)\; p(c|I) \qquad . \tag{12}$$

In order to facilitate analytic calculation the exponent of the likelihood is written in a quadratic form over $c$

$$\sum_i [d_i - c f_i]^2 = (\vec{f}^T\vec{f})[c-c_0]^2 + \left[\vec{d}^T\vec{d} - \frac{(\vec{d}^T\vec{f})^2}{\vec{f}^T\vec{f}}\right] \qquad , \tag{13}$$

where $c_0 = \vec{d}^T\vec{f}/\vec{f}^T\vec{f}$. For the prior $p(c|I)$ the only thing we know is that $c$ will be something in between an upper and a lower limit, where it is reasonable to assume that the upper (lower) bound is given by an unknown factor $n$ $(1/n)$ of the value $c_0$ where the maximum of the likelihood occurs. The principle of maximum entropy gives a flat prior with

$$p(c|I) = \begin{cases} \frac{1}{nc_0} & \forall\ 0 \le c \le nc_0 \\ 0 & \text{else} \end{cases} \qquad . \tag{14}$$

The integral over the c-dependent parts then reads

$$\frac{1}{nc_0}\int_0^{nc_0} \mathrm{d}c \;\exp\left\{-\frac{1}{2\sigma^2}(\vec{f}^T\vec{f})[c-c_0]^2\right\} \qquad . \tag{15}$$

One may check that for $\vec{f}^T\vec{f} \gg \sigma^2$ it is allowed to shift the integral boundaries to $+/-$ infinity with affecting the value of the integrand up to a small error only. As a matter of fact for the chosen model parameters of $N$=3, $x_i$=$i$, $T$=1, c=1 and $\sigma$=0.1 the error is in the order of $10^{-7}$ of the correct integral. Notice that this is almost the same for every $T$ in between 0 and infinity. We finally get

$$p(T|D,I) \propto \frac{\sqrt{\vec{f}^T\vec{f}}}{\vec{d}^T\vec{f}} \exp\left\{-\frac{1}{2\sigma^2}\left[\vec{d}^T\vec{d} - \frac{(\vec{d}^T\vec{f})^2}{\vec{f}^T\vec{f}}\right]\right\} \qquad . \tag{16}$$

A look at the behavior for large and small T gives

$$\begin{aligned} \lim_{T\to 0} p(T|D,I) &\propto \frac{\sqrt{N}}{\sum_i d_i} \exp\left\{-\frac{1}{2\sigma^2}\left[\vec{d}^T\vec{d} - \frac{(\sum_i d_i)^2}{N}\right]\right\} \\ &\propto const \qquad , \end{aligned} \tag{17}$$

$$\begin{aligned} \lim_{T\to \infty} p(T|D,I) &\propto \frac{\sqrt{\vec{x}^T\vec{x}}}{\vec{d}^T\vec{x}} \exp\left\{-\frac{1}{2\sigma^2}\left[\vec{d}^T\vec{d} - \frac{(\vec{d}^T\vec{x})^2}{\vec{x}^T\vec{x}}\right]\right\} \\ &\propto const \qquad . \end{aligned} \tag{18}$$

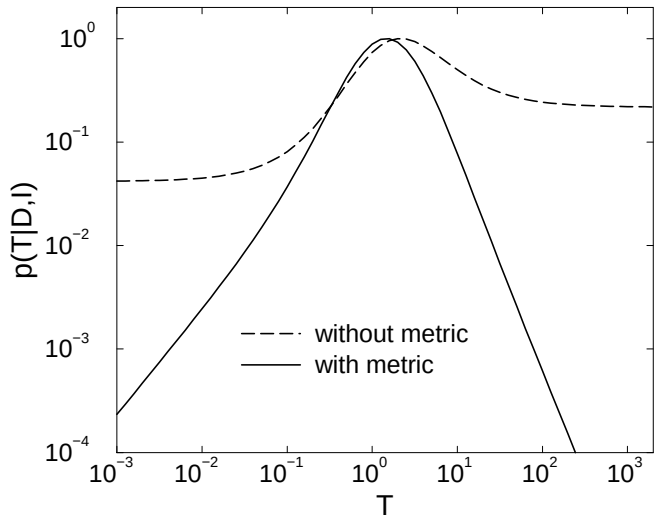


FIG. 1: Posterior $p(T|D,I)$ with (solid line) and without (dashed line) the Riemannian metric. Neglection produces non-vanishing tails.

Though one has no problem with the lower limit since the integrand is regular, the non-vanishing posterior distribution for $T \to \infty$ leads to an expectation value which depends on where the integration limits are set (see Fig. 1).

Now we implement in the calculation the Riemannian metric. From Eq. (10) we get an additional factor $c$, so the integration over the $c$-dependent parts changes to

$$\frac{1}{nc_0}\int_0^{nc_0} \mathrm{d}c\ c\ \exp\left\{-\frac{1}{2\sigma^2}(\vec{f}^T\vec{f})[c-c_0]^2\right\} \quad . \tag{19}$$

Again it is allowed to extend the integration limits to $+/-$ infinity with only minor error. The full posterior then gives

$$p(T\,|D,I) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[\vec{d}^T\vec{d} - \frac{(\vec{d}^T\vec{f})^2}{\vec{f}^T\vec{f}}\right]\right\}\sqrt{\left(\frac{\partial\vec{f}}{\partial T}^T\frac{\partial\vec{f}}{\partial T}\right) - \left(\frac{\partial\vec{f}}{\partial T}^T\vec{f}\right)^2/(\vec{f}^T\vec{f})}\ . \tag{20}$$

What is now the behavior of $p(T|D,I)$ for $T$ approaching 0 and infinity? The exponent in Eq. (20) was already examined in Eqn. (17) and (18) to become constant, so we only have to look at the square root.

$$\lim_{T\to 0}\sqrt{\sum_i\left(\frac{x_i}{T+x_i}\right)^4 - \left[\sum_i\left(\frac{x_i}{T+x_i}\right)^3\right]^2/\sum_i\left(\frac{x_i}{T+x_i}\right)^2} = \sqrt{N-\frac{N^2}{N}} = 0\ , \tag{21}$$

$$\lim_{T\to\infty}\frac{1}{T^2}\sqrt{\sum_i\left(\frac{x_i}{1+x_i/T}\right)^4 - \left[\sum_i\left(\frac{x_i}{1+x_i/T}\right)^3\right]^2/\sum_i\left(\frac{x_i}{1+x_i/T}\right)^2} = 0 \quad . \tag{22}$$

So indeed the square root term which stems from the metric does take care of zero tails in the posterior! The nice decrease towards 0 is shown in Fig. 1 by the solid line.

## IV. REAL WORLD PROBLEM

In the problem of determining the electron temperature in an hydrogen plasma heated by electron cyclotron resonance, the model function $T$ depends in a quite complicated way on the temperature $T$:

$$\vec{f}(T) = -\boldsymbol{V}(\boldsymbol{R}-\boldsymbol{V})^{-1}\vec{x} \qquad . \tag{23}$$

Both $\boldsymbol{V}$ and $\boldsymbol{R}$ are matrices, but only the diagonal matrix $\boldsymbol{V}$ depends on T with entries on the diagonal:

$$V_{ii} = \frac{1}{\frac{a_i}{\sqrt{T}} + \frac{1}{b_i T}} \qquad , \tag{24}$$

where $a_i$ and $b_i$ are constants with respect to ion species $i$. Since the sensitivity of the measurement apparatus is unknown one has to introduce a linear parameter $c$ in order to relate the data to the model, i.e. Eq. (1). Contrary to our simple problem we are not so fortunate to know the variance $\sigma$ exactly. The experimentalist can only provide an estimate $\vec{s}$ of the true errors $\vec{\sigma}$ with respect to each other but not on the total scale, so that we have to introduce an overall multiplication factor $\omega$, with $\sigma_i = \omega s_i$. In order to assign a prior to $\omega$ the outlier tolerant approach [4] was chosen:

$$p(\omega|\alpha,\gamma I) = 2\frac{\alpha^\gamma}{\Gamma(\gamma)}\left(\frac{1}{\omega}\right)^{2\gamma}\exp\left\{-\frac{\alpha}{\omega^2}\right\}\frac{1}{\omega} \qquad . \tag{25}$$

The expectation value of $\omega$ should be one, since the experimentalist does his estimation according to his best knowledge. Furthermore, from the characteristics of the measurement process one can tell that the best guess of $\vec{s}$ should not deviate by more than 50% from the true $\vec{\sigma}$. This results in $\alpha = 1.28$ and $\gamma = 2.0076$.

Now we follow the route explained above to evaluate the expectation value of $T$. Again we start by marginalizing $c$ (with the flat prior of Eq. (14)) and $\omega$ without making use of the Riemannian metric. This gives the posterior in $T$

$$p(T|D,I) \propto \frac{\sqrt{\vec{\hat{f}}^T\vec{\hat{f}}}}{\vec{\hat{d}}^T\vec{\hat{f}}}\left[\alpha + \frac{1}{2}\left(\vec{\hat{d}}^T\vec{\hat{d}} - \frac{(\vec{\hat{d}}^T\vec{\hat{f}})^2}{\vec{\hat{f}}^T\vec{\hat{f}}}\right)\right]^{-\frac{N}{2}-\gamma-1} \qquad . \tag{26}$$

For simplicity of notation the hat shall denote that the values have been divided by the

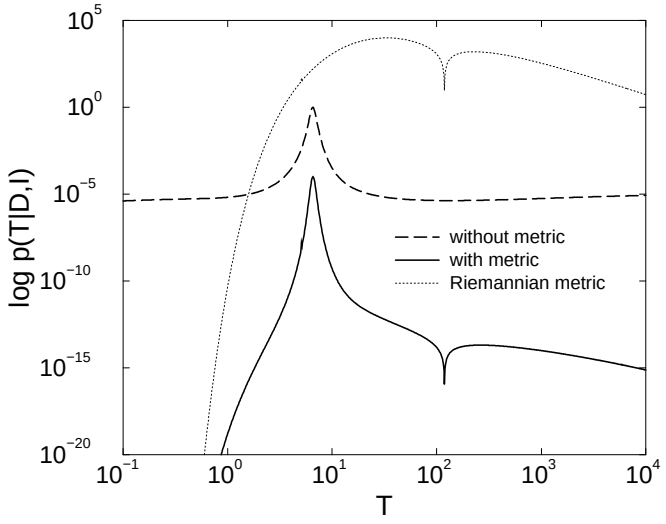


FIG. 2: Posterior $p(T|D,I)$ with (solid line) and without (dashed line) the Riemannian metric (dotted line). The incision at T = 118.43 K is a single point which is due to the parameterization of the physical model. It does not affect the integrability.

estimated error $\vec{s}$: $\hat{d}_i = d_i/s_i$. The posterior is displayed in Fig. 2. Here we have to face the problem we observed above in the simple example. Though a non-vanishing tail for $T \to 0$ is not so harmful, the increase with $T \to \infty$ results in a divergence.

Help comes by obeying the correct Riemannian metric. Then the posterior reads

$$p(T|D,I) \propto \mu(T) \frac{1}{\sqrt{\vec{\hat{f}}^T \vec{\hat{f}}}} \left[ \alpha + \frac{1}{2} \left( \vec{\hat{d}}^T \vec{\hat{d}} - \frac{(\vec{\hat{d}}^T \vec{\hat{f}})^2}{\vec{\hat{f}}^T \vec{\hat{f}}} \right) \right]^{-\frac{N}{2}-\gamma-1} \tag{27}$$

where $\mu(T)$ is just the metric of Eq. (10) without the terms in $c$ and $\omega$ (marginalized over). The situation changes completely (see Fig. 2) and the integral becomes feasible now.

## V. CONCLUSION

The correct mathematical way to deal with marginalization integrals is to use the Riemannian metric. This invariant measure takes care of defining correct infinitesimal elements to be integrated over. Since parameterizations of a model may be subjective and vary with the investigator of a problem, this is the only consistent way to get comparable answers in probability space.

## VI. ACKNOWLEDGMENT

We like to acknowledge discussions with C. Rodriguez.

---

[1] C. Rodriguez, "The metrics induce by the kullback number," in *Maximum Entropy and Bayesian Methods*, J. Skilling, ed., Kluwer Academic, Dordrecht, 1989.

[2] S. Amari, *Differential-Geometrical Methods in Statistics*, Springer-Verlag, Berlin, Heidelberg, 1985.

[3] C. Rodriguez, "From euclid to entropy," in *Maximum Entropy and Bayesian Methods*, J. W. T. Grandy, ed., Kluwer Academic, Dordrecht, 1991.

[4] V. Dose and W. von der Linden, "Outlier tolerant parameter estimation," in *Maximum Entropy and Bayesian Methods*, V. Dose et al., ed., Kluwer Academic, Dordrecht, 1999.